Political Science 15
Introduction to Research in Political Science
Lecture 4a: Introduction to Bivariate Regression

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Recap: what we've seen so far

Lecture 3: Probability

- How to read a probability density function
- Key terms: random variables, expectation, variance
- Difference between population and sample
- 2 (powerful) theorems: Law of Large Numbers and Central Limit Theorem
- Distribution of the sample mean $\bar{X}$:
  1. Centered around $\mathbb{E}[X]$
  2. Variance of $\bar{X}$ is $var(X)/n$
  3. Normal distribution

Housekeeping

Problem Set 1

- Your grades have been posted
- I will post a review video
- It is your responsibility to ask your TAs where you lost points (in office hours today)

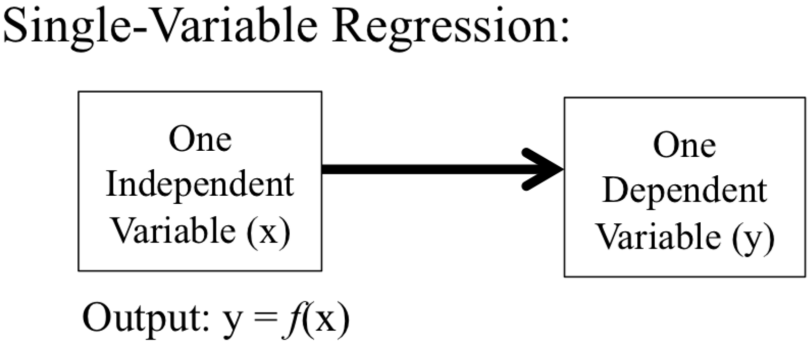
Lecture modules

- ∗ If you are watching the lecture modules by downloading the videos, email me so that I can give you credit for participation! (Should only apply to 1 or 2 of you.)

Motivation

- We want to know about relationships in the world: does smoking cause cancer? do vitamins increase lifespan? does democracy cause economic development? does HCL cure COVID-19?
- First, we talked about how, even though we see a correlation between two variables, it's not necessarily a causal relationship.
- Next, we talked about how exogeneity, usually through random assignment of the treatment, allows us to make a causal claim.
- Then, we discussed the building blocks for probability and uncertainty. This will help us discern when an observed relationship is really there, or if it is just by chance (more later).
- Today, we will continue, discussing how we can formally detect relationships through linear regression.
- It's useful to remember the equation $Y = mX + b$ from high school math. This is very similar to a regression.

Set-up of a bivariate regression



Single-Variable Regression:

Output: y = *f*(x)

Bivariate regression equation (from Lecture 2e)

Y_i is the **dependent variable**, or outcome of interest

X_i is the **independent variable**, i.e. a treatment or explanatory factor

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

ϵ_i is the **error term**, or everything that we haven't captured in our model. We cannot observe this!

β_0 is the **intercept/constant**, or the value of Y when X is 0

β_1 is the **slope**, or how much change in Y is associated with a one-unit change in X

Relationships between variables

So far we have talked about one variable at a time (how it is distributed), or about how one variable looks across two groups, e.g.:

- election turnout for young and old people
- treated and control group
- distribution of income (typically right-skewed)

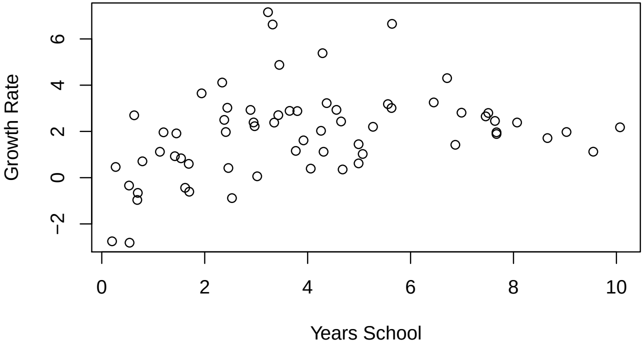
Suppose you are interested in relationships between two variables and neither is really binary, for example:

- years of war with high versus low ethnic fractionalization (Problem Set 2)
- re-election rate for incumbent presidential parties and economic performance (textbook example)
- economic growth and average education level across countries
- economic development and natural resources

Visualization: scatter plot

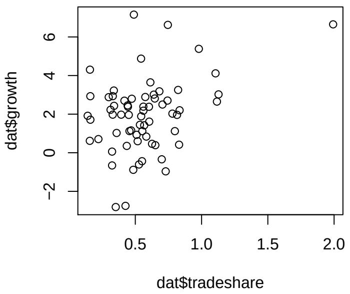
Let's look at average adult education (horizontal axis) and growth rate in real GDP (vertical axis):

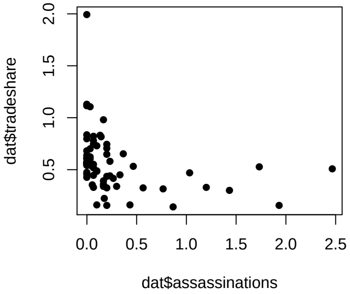
```
plot(dat$yearsschool, dat$growth,
xlab = "Years School", ylab = "Growth Rate")
```



Let's see some more:

```
>names(dat)
[1] "country_name"   "growth"         "rgdp60"
[4] "tradeshare"     "yearsschool"    "assassinations"
> plot(dat$tradeshare, dat$growth)
> plot(dat$assassinations, dat$tradeshare, pch=16)
```





Next time

In the next module, we will see how to **quantify** the association between X and Y, rather than just eyeballing it.

Political Science 15
Introduction to Research in Political Science
Lecture 4b: Measures of Association

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

How to measure association

We want to measure the association between two random variables X and Y.

3 options to measure association

1. Covariance
2. Correlation
3. Bivariate regression

Option 1: Covariance

The first measure of association we will talk about is covariance. (See Appendix in *Real Stats*).

It will help us determine whether two variables are:

- **positively associated**: when X is higher, we expect Y to be usually higher
- **negatively associated**: when X is higher, we expect Y to be usually lower
- **not associated**: when X is higher, it doesn't tell us anything about Y

Covariance

As the name suggests, it is analogous to variance but associates two variables (co-) and describes *how the two vary together*.

The **population covariance** is a function of the distributions:

- Variance: $Var(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$
- Covariance: $Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$

In an observed **sample** (not the population), we use analogous estimators:

- Sample variance: $\widehat{Var}(X) = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(X_i - \bar{X})^2$
- Sample covariance:

$$\boxed{\widehat{Cov}(X, Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}$$

Problems with Covariance

Problem with covariance: the scale is not very natural. **Covariance is hard to interpret!**

How to interpret covariance:

- If the covariance is positive, I can tell you that as X increases, I expect Y to increase.
- If the covariance is negative, I can tell you that as X increases, I expect Y to decrease.
- If the covariance is 0, I can tell you that knowing about X will teach me nothing about Y.
- That's it!

While covariance can tell us about the *direction* of the association (positive/negative), it cannot tell us about the *magnitude* of the association (strong/weak association).

Option 2: Correlation

Wouldn't it be easier if it went between -1 and $+1$?

Correlation is a rescaled version of covariance.

How to interpret correlation

- a perfect positive relationship has $cor(X, Y) = 1$
- a perfect negative relationship has $cor(X, Y) = -1$
- two perfectly unrelated variables have $cor(X, Y) = 0$

We get this by dividing the covariance by the product of the standard deviations of X and Y, which effectively standardizes/rescales the covariance:

$$\boxed{cor(X, Y) = \frac{cov(X, Y)}{SD(X)SD(Y)}}$$

The correlation is often denoted $\rho(X, Y)$ ("rho") or simply r.

Option 2: sample estimator of correlation

The previous equation was the **population** version of correlation. We want the **sample** version, so we use the following estimator:

compare to sample covariance

$$\hat{\rho}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

compare to standard deviation of X
compare to standard deviation of Y

Note: the $\frac{1}{n-1}$ cancel out!

Correlation in R

One way to calculate the correlation is to divide the covariance `cov()` by the product of the standard deviations `sd()`.

```
> cov(dat$growth,dat$yearsschool)/(sd(dat$growth)*sd(dat$yearsschool))
[1] 0.3309986
```

But R has a built-in function `cor()` that you can use instead!

```
> cor(dat$growth, dat$yearsschool)
[1] 0.3309986
```

What do you get for $cor(X, X)$?

Check: what do you get for $cor(X, X)$?

You should be able to guess given the meaning of $cor()$.

The answer is 1. But let's verify this:

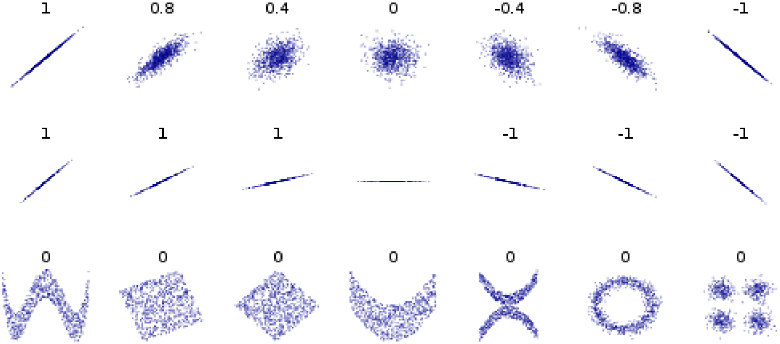
$$\begin{aligned}
\rho(X, X) &= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]}{\sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}\sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]}} && \text{(correlation equation)} \\
&= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])\right]}{\sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}\sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}} && \text{(plug in } X\text{)} \\
&= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{\mathbb{E}[(X - \mathbb{E}[X])^2]} && \text{(everything cancels out)} \\
&= 1
\end{aligned}$$

$\Rightarrow$ X is perfectly positively correlated with itself!

Warning: Correlation only sees “linear” relationships

Imagine you observe samples that give you the following clouds of data points. The number at the top is the correlation.

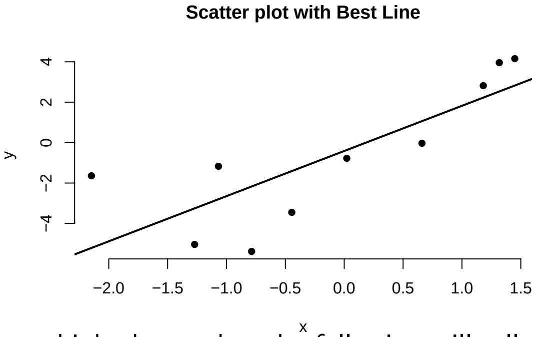
- Correlation does a good job at capturing linear relationships (top 2 rows).
- But it does a poor job at capturing non-linear relationships (bottom row).



Correlation does *not* capture all meaningful relationships between X and Y.

Option 3: Bivariate Regression

Imagine placing a “best-fitting” line on the cloud of data points. We will define what we mean by “best” next time.



In the meantime, think about what the following will tell you about the association between X and Y:

- Sign of the slope? Direction of the association between X and Y.
- Magnitude of the slope? Strength of the association between X and Y.

Option 3: Bivariate Regression

Recall the bivariate regression equation (AKA most important equation of the course) which posits a linear relationship between X and Y.

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i}$$

β_0 and β_1 tell us very specific things:

- β_0 tells us the value of Y when $X = 0$
- β_1 tells us by how much Y will change when X increases by 1 unit

As we will see, bivariate regression is a very useful measure of association. It is the workhorse of statistics.

Turns out, the slope of this equation is:

$$\hat{\beta}_1 = \frac{cov(X, Y)}{Var(X)}$$

This is the same as Equation 3.4 in the textbook, just written differently.

Next time

More about bivariate regression, including:

- What exactly do we mean by “best-fitting” line?
- How can we use it to make predictions from our data?
- How do we interpret a regression table?

Political Science 15
Introduction to Research in Political Science
Lecture 4c: Regression in Practice

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Recap of Lecture 4 so far

In the previous modules we talked about:

- Measures of association between two variables
    1. Covariance
    2. Correlation
    3. Bivariate regression
- Bivariate regression places the “best-fitting” line on a scatter plot

Today:

- What exactly do we mean by “best-fitting” line?
- How can we use bivariate regression to make predictions from our data?
- How do we interpret a regression table?

Hint: Regression analysis is sometimes called

OLS = Ordinary Least Squares

Regression logic

We want to find the line that best describes the relationship between X and Y, i.e. that best ‘fits’ the data.

That is,

- We want to make a line that best shows the relationship between X and Y.
- So, we want to predict Y ($\hat{Y}$) using our slope (β_1) and intercept (β_0) from X. We want to make the best prediction we can!
- Think of this $\hat{Y}_i$ as a prediction for Y_i that takes X_i into account.
- If X is a certain value, what would we expect Y to be?

How does OLS place the best-fitting line?

- How can we best figure out this relationship between X and Y?
- Suppose we guess β_0 and β_1. How do we assess whether it is a good guess for the relationship between X & Y?

- We use the sum of squared errors (the textbook uses residuals) to see how well we are doing:

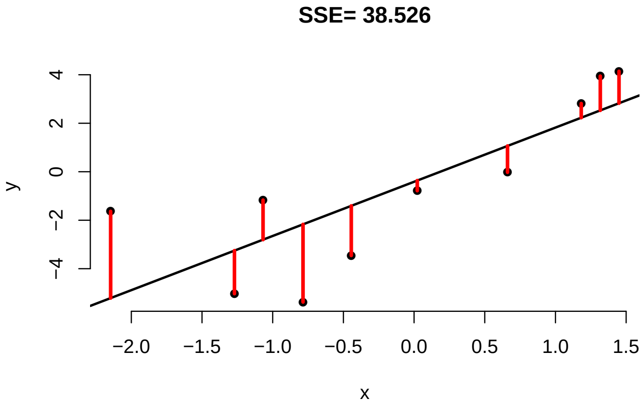
$$SSE = \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- It turns out the best way to estimate this relationship is to choose our slope and intercept (β_0 and β_1) to **minimize the sum of squared errors** (SSE).
  (*Note: also called the sum of squared residuals (SSR) or the residual sum of squares (RSS)*).

Regression under the hood

- Peek the math: calculus helps us estimate the **parameters** β_0 and β_1 that minimize the SSE.
- The choice of $\beta_1 = \frac{Cov(X,Y)}{Var(X)}$ makes this SSE smallest, so we say it is "best-fitting line in least squares sense".

Best-fitting line

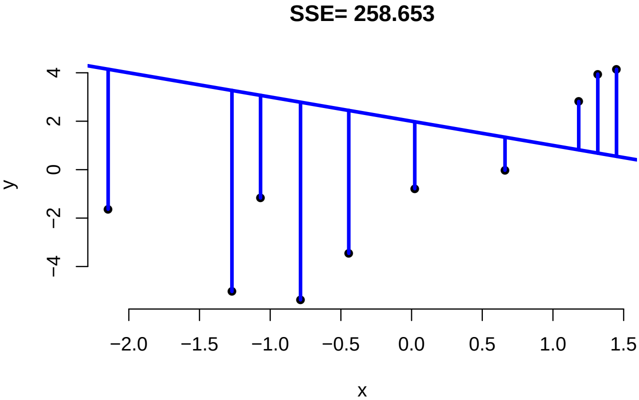


The residual is the distance between the line and any given point.

The SSE takes those residuals, squares them, and adds them up.

The regression shows us the best fitting line in terms of sum of squared errors (SSE).

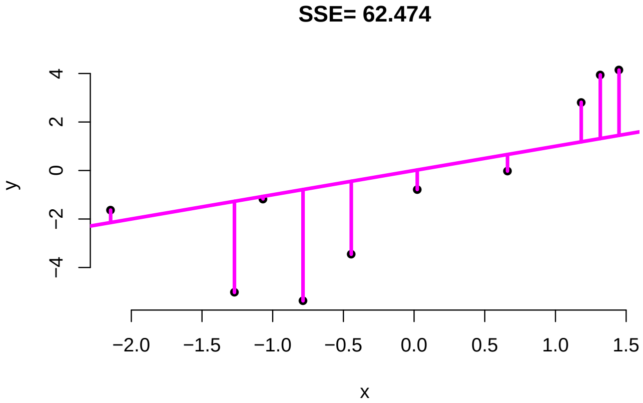
Compare to this



This is a pretty terrible fit.

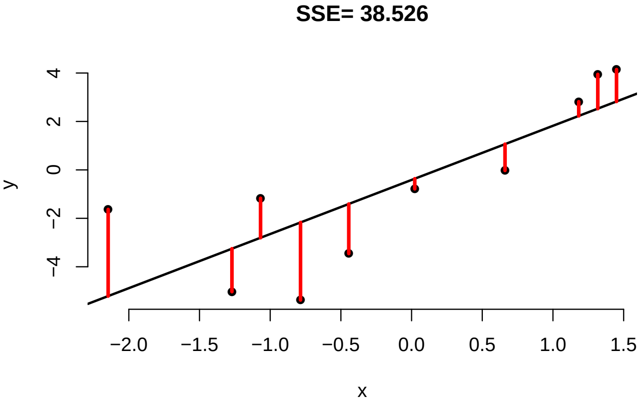
Sum of squared errors (SSE) is larger: 259 compared to 39.

And this



This is not totally unreasonable, but could be improved.
SSE = 62, not as bad as last one, but not as good as 39.

Let's go back to the best fit



For this best fitting line, $\hat{\beta}_0 = -.41$, $\hat{\beta}_1 = Cov(x, y)/Var(x) = 2.24$.

Estimating the regression model

Recall the canonical model for bivariate OLS:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- X is the *independent variable* or *predictor* or *covariate*, and
- Y is the *dependent variable*
- β_0 is the *intercept*: the value of Y_i we will guess at $X = 0$.
- β_1 is the *slope*: a one-unit change in X is associated with a β_1 change in Y
- ϵ_i called the *error*

Then we use data to estimate this model:

- We run our estimator (bivariate OLS) to get estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$
- **Fitted/predicted value**: $\hat{Y}_i$ is our best guess of Y_i for individual i with characteristic X_i

  $\Rightarrow$ Construct it with: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- **Fitted residual**:, $Y_i - \hat{Y}_i = \hat{\epsilon}_i$

Regression in R

For variables saved as x and y you could estimate β_1 by the formula, $\hat{\beta}_1 = \widehat{Cov}(x, y)/\widehat{Var}(x)$:

```
> beta1=cov(x,y)/var(x)
> beta1
[1] 2.235034
```

To get $\hat{\beta}_0$, you want mean of fitted residuals to equal 0, and you get that when $\hat{\beta}_0 = mean(Y_i - \hat{\beta}_1 X_i)$:

```
> beta0=mean(y-beta1*x)
> beta0
[1] -0.4143977
```

Regression in R

Typically though, we use an R function called `lm()`, for “linear model”:

```
> lm.out = lm(y˜x)
> summary(lm.out)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4144     0.6969  -0.595  0.56853
x             2.2350     0.5922   3.774  0.00544 **
---
```

Note: Just like the equation for a regression, Y goes first. In other functions in R, like `plot()`, X goes first. Think about the order in which you type things.

The intercept (β_0) and slope (β_1) are called **coefficients**, and listed under “Estimate” column.

Your first regression table!

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4144     0.6969  -0.595  0.56853
x             2.2350     0.5922   3.774  0.00544 **
```

This table with show you the estimate for your two coefficients – the value on your IV and your intercept.

Let's practice reading this table and interpreting these two values.
$\hat{\beta}_0$: When X is zero, we expect Y to equal -0.4144.
$\hat{\beta}_1$: For each one-unit change in X, we expect a 2.2350 unit change in Y.

You also get something else: a standard error for each coefficient!

- If we took another sample of data from the same source and estimated $\hat{\beta}$ again, you would get different $\hat{\beta}$. Think of the distribution of such βs.
- $SE(\hat{\beta})$ estimates the standard deviation for that distribution of $\hat{\beta}$.

A more interesting example

Enough of x and y, let's go back to *yearsschool* and *growth*

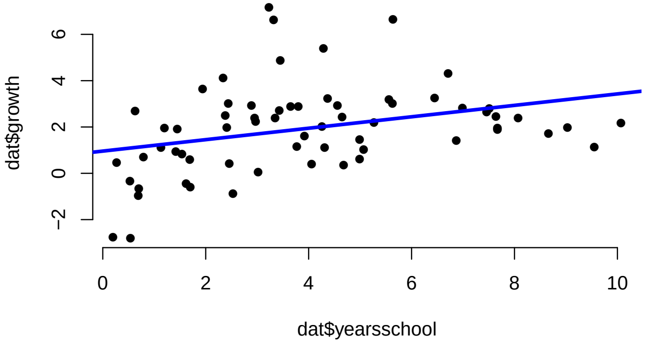
- We'll call *yearsschool* the independent variable/predictor, and *growth* the dependent variable/outcome.
- We say we "regress (outcome) on (predictor)", so we "regress *growth* on *yearsschool*".

```
> lm.out=lm(dat$growth~dat$yearsschool)
> ## Or, same thing this way:
> lm.out=lm(growth~yearsschool, data=dat)
> summary(lm.out)
Call:
lm(formula = growth ~ yearsschool, data = dat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95829    0.41846   2.290  0.02538 *
yearsschool  0.24703    0.08873   2.784  0.00708 **
---
Residual standard error: 1.804 on 63 degrees of freedom
Multiple R-squared:  0.1096,Adjusted R-squared:  0.09543
F-statistic: 7.752 on 1 and 63 DF,  p-value: 0.007077
```

What your regression is saying

```
plot(dat$yearsschool, dat$growth, pch=16, frame.plot=FALSE)
abline(lm.out, lwd=3, col=4)
```



Practice and more interpretation of the regression table

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95829    0.41846   2.290  0.02538 *
yearsschool  0.24703    0.08873   2.784  0.00708 **
---
Residual standard error: 1.804 on 63 degrees of freedom
Multiple R-squared:  0.1096,Adjusted R-squared:  0.09543
F-statistic: 7.752 on 1 and 63 DF,  p-value: 0.007077
```

- How to interpret the intercept?
  At *yearsschool* = 0, we expect growth of 0.96.

- How to interpret the *yearschool* coefficient estimate? A one year change in *yearsschool* is associated with a 0.25 unit change in expected *growth*.

- Note that the units in the above statement reflect the original units of the X and the Y.

- The "R-squared" tells us how much of the variation of Y (*growth*) is "explained" by X (*yearsschool*).

Next time

- Diagnostics to see how well our model performs
- Properties of OLS
  1. Bias
  2. Consistency
- Problems with OLS
  1. Heteroskedasticity
  2. Outliers. **Read Wand et al. 2001!**

Political Science 15
Introduction to Research in Political Science
Lecture 4d: Properties of OLS

Alice Lépissier
University of California Santa Barbara

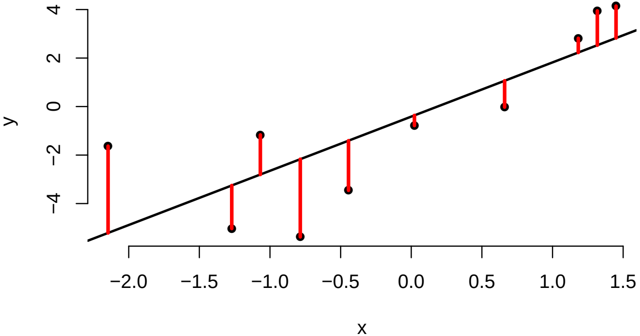
Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Recap of Lecture 4 so far

In the previous modules we talked about:

- How to estimate bivariate regression
  - Line of best fit minimizes the Sum of Squared Errors
  - OLS estimates the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the SSE
- How to interpret a regression table
- Fitted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and residuals $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

SSE= 38.526



Lecture 4

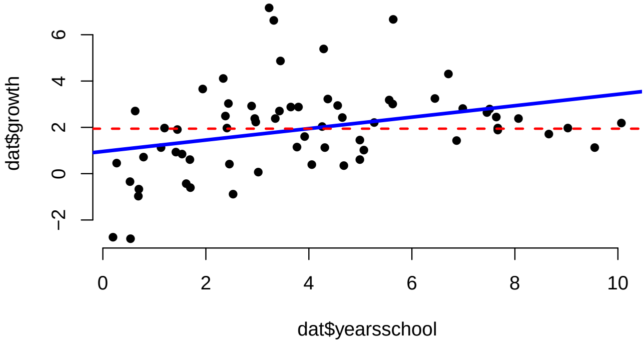
Today:

- Diagnostics: how to assess model “fit” (or performance)
- Bivariate regression and causal inference
- Properties of OLS
  - Bias
  - Consistency

Compare modeled outcome to the simple mean

We can think of regression as a prediction machine that tells us our best guess of Y (*growth*) given our knowledge of X (*yearsschool*) for an observation.

So let's compare what we guess using regression to what we'd guess if we only knew the mean of Y (*growth*) and did not have information from X (*yearsschool*).



Understanding the “Variance Explained” or R^2

The “spread” of points around the regression line should be smaller than the spread of points around the (red) mean-line.

- If we average up the squared distances around the mean line we get the variance of Y. Here, $var(growth) = 3.60$.
- If we average up the squared distances around the regression line we get mean squared error (MSE) for the regression:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - (\beta_0 + \beta_1 X_i))^2$$

What does the MSE equation remind you of? This is the average of the SSE equation!

Understanding the “Variance Explained” or R^2

- We want to minimize this form of prediction error in our choice of β.

```
> yhat=predict(lm.out)
> mean((dat$growth-yhat)^2)
[1] 3.155446
```

You could say $\dfrac{MSE}{Var(Y)}$ is the proportion of variance of Y that remains unexplained by the model. Here, it is $3.16/3.60 = 0.88$

Thus $1 - \dfrac{MSE}{Var(Y)} = 0.12$ is the proportion of variance “explained”

- this is R^2 statistic at the bottom of the table
- it is also equal to the square of $cor(X, Y)$!

Causal Inference

Does this mean *yearsschool* is a *cause* of *growth*?

NO!!!!

Just like difference in means reflected an association, correlations, covariances, and regression coefficients only reflect an observed relationship in the data.

- the setup makes us focus on variation in *yearsschool* as an explanation for *growth*
- but those countries that are higher or lower on *yearsschool* are probably higher or lower on lots of other things
- these other things may be why *growth* differs, not the *yearsschool*
- in terms of confounders: something may cause some countries to have higher education and may also cause those countries to have higher growth.
- what are some examples of such possible confounders here?

Key Formulae

Regression

For the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The best-fitting line in the least-squares sense minimizes the sum of squared errors, or equivalently the mean squared error (MSE),

$$MSE = \frac{1}{N} \sum_i \left[(Y_i - (\beta_0 + \beta_1 X_i))^2 \right]$$

The sample estimate for its coefficients are:

- $\hat{\beta}_1 = \frac{\widehat{cov}(X, Y)}{\widehat{Var}(X)}$
- $\hat{\beta}_0 = mean(Y_i - \hat{\beta}_1 X_i)) = \overline{Y} - \hat{\beta}_1 \overline{X}$

The variance explained is $\hat{R}^2 = 1 - \frac{MSE}{\widehat{Var}(Y)}$

Review of key points on OLS so far...

- We can estimate the relationship between X & Y, with two key coefficients: $\hat{\beta}_1$ & $\hat{\beta}_0$. Check: why are there hats?
- Due to sampling variation, our estimates also have a standard error
- Our model creates fitted values, $\hat{Y}$
- Our model has residuals, the part of Y that X does not explain, $\hat{\epsilon}_i$
- Relatedly, we can estimate our model's 'goodness of fit': R^2
- Just because we see a relationship in our regression model, doesn't mean its causal

More key points on OLS

- Bias in estimates
- Consistency of estimates

Pro-tip: What are two nice ways to compliment an estimator?
Call it unbiased and consistent.

We will review all these ideas in this set of slides, and you can find more explanation in the textbook.

Bias in Estimators

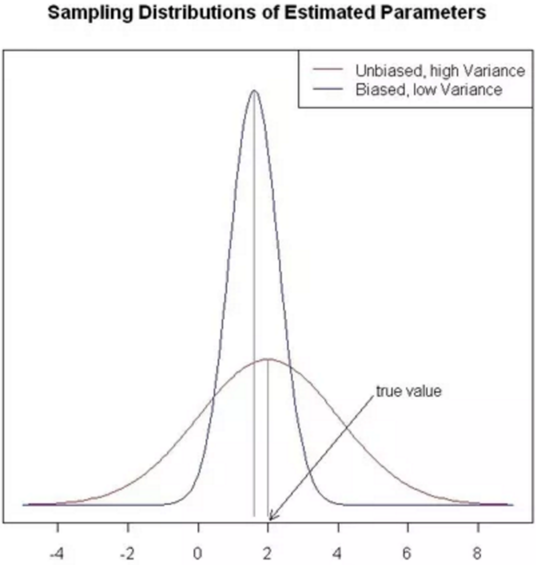
- What is **bias**?
  - Unbiased estimate: On *average*, our estimate is equal to the true parameter, $\hat{\beta}_1 = \beta_1$
  - Biased estimate: Our coefficient is systematically wrong, either too high or two low when compared to the true parameter, $\hat{\beta}_1 > \beta_1$

Omitted Variable Bias

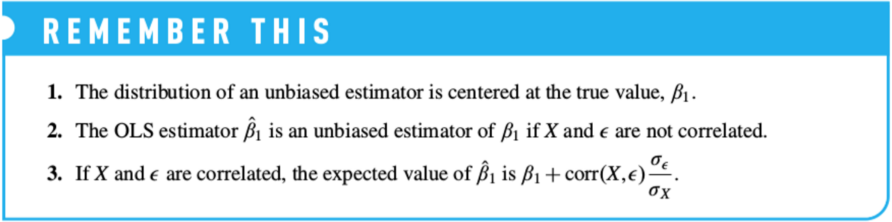
- OLS does *not* necessarily create unbiased estimates. Why?

- **Omitted variable bias**: this is a specific form of endogeneity.
  - X is correlated with something else that influences Y; the error term is correlated with Y.
  - This is often the reason why, if you change the model specification, your estimates change. In this case, we say that our model is not *robust* to alternative specifications.
  - Therefore, our model is missing a key confounder, and we haven't estimated a causal relationship.
  - Theoretically, you could include this confounder in your model (multi-variable regression) but this is often hard to do: maybe you don't know what the confounder is, or you don't have data on it.

Bias-Variance trade-off



Key Points on Bias + OLS



REMEMBER THIS

1. The distribution of an unbiased estimator is centered at the true value, β_1.
2. The OLS estimator $\hat{\beta}_1$ is an unbiased estimator of β_1 if X and ϵ are not correlated.
3. If X and ϵ are correlated, the expected value of $\hat{\beta}_1$ is $\beta_1 + \text{corr}(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$.

So, how do we get around this bias? Exogeneity!

Consistency in Estimates

- Consistency:

  As our sample size (N) grows, the estimate converges on the true parameter:

$$\text{plim } \hat{\beta}_1 = \beta_1$$

- OLS estimators are consistent.
- But remember, your estimate will only converge on the true parameter (be consistent) if there is no omitted variable bias in your regression model.

A Picture of Consistency

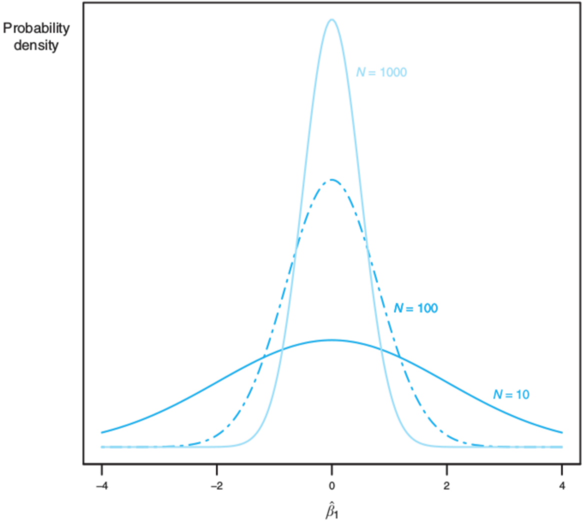


FIGURE 3.8: Distributions of $\hat{\beta}_1$ for Different Sample Sizes

Political Science 15
Introduction to Research in Political Science
Lecture 4e: Problems with OLS

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Recap of Lecture 4

Last module, we discussed key properties of the OLS estimator:

- Unbiasedness (if and only if the independent variable X is uncorrelated with the error term ϵ)
- Consistency

Today, we discuss several problems that may arise with OLS and what to do in that case.

- Heteroskedasticity
- Outliers

Reminder

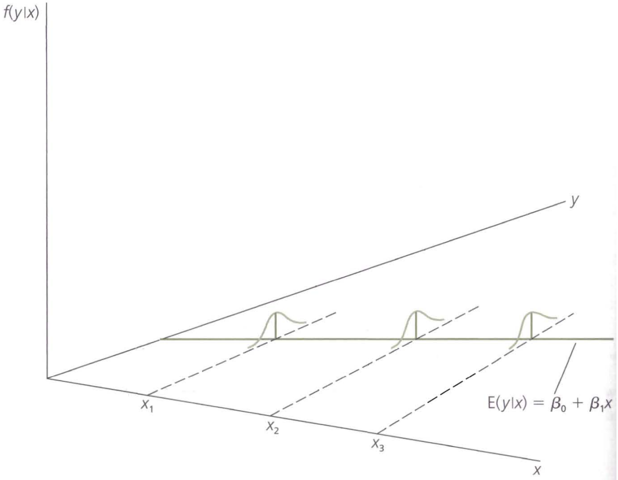
You need to have read the Wand et al. 2001 paper on "The Butterfly Did It" for the section on outliers to make sense.

Homoskedasticity vs Heteroskedasticity

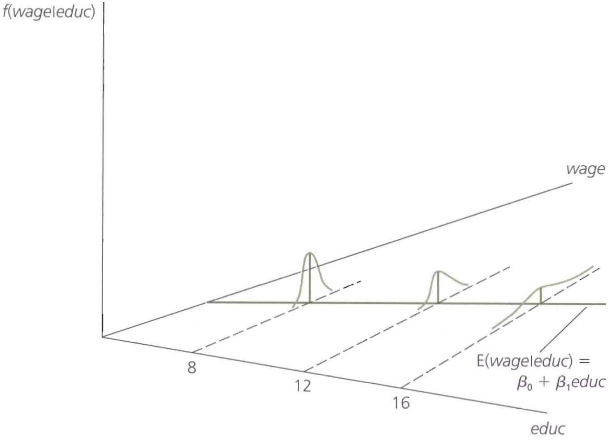
- Rather than worrying about the bias of our $\hat{\beta}_1$ estimates, we could worry about the bias of our σ^2 estimates.
- Homoskedasticity: When the error term, ϵ, has the same variance for all observations of X. This isn't a problem.
- Heteroskedasticity: When the error term, ϵ, **does not** have the same variance for all observations of X. This is a fixable problem.

- When we have heteroskedasticity, it means we have a non-constant variance in our errors.
- It's easy to see this with some pictures.

A Picture of Homoskedasticity



A Picture of Heteroskedasticity



Fixing the Heteroskedasticity Problem

- Heteroskedasticity: When the error term, ϵ, **does not** have the same variance for all observations of X. This is a fixable problem.

- Remember, this only affects our standard errors, $var(\hat{\beta}_1)$, not our $\hat{\beta}_1$ estimate. Therefore, it doesn't cause bias.

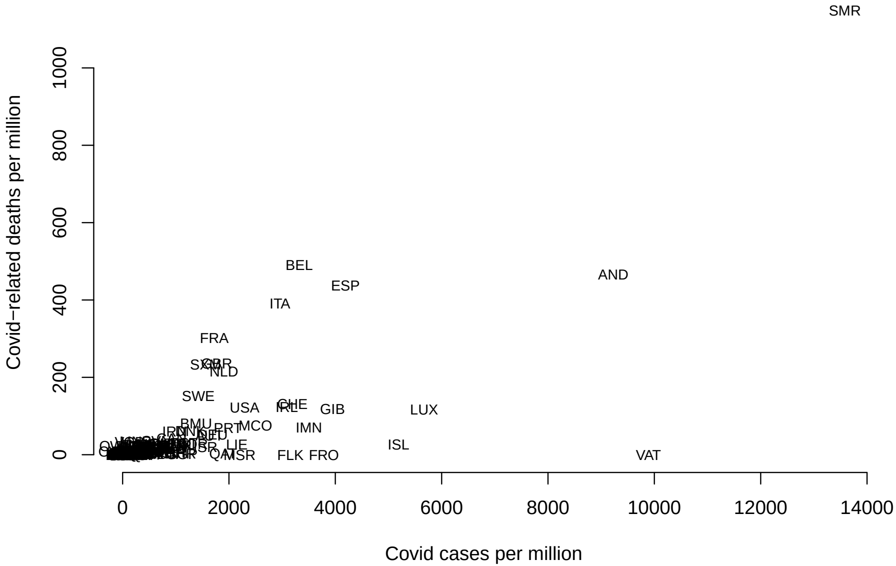
- So what's the solution? We use a slightly different estimator for our standard error calculations (e.g. White's heteroskedasticity-consistent estimator). Intuitively, we estimate the variance in our standard errors.

- You can do this in R. (See computing corner).

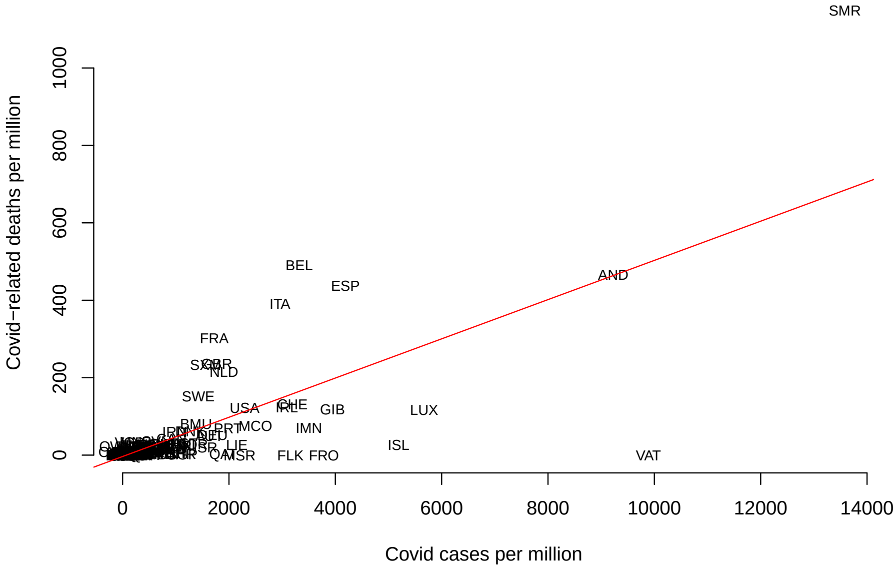
Outliers

- Reminder, what is an outlier? An observation that is extremely different from the rest of the observations in the sample. “One of these things is not like the others.”
- Reminder, what does an outlier do to our estimate of the mean? It drags it towards the outlier. The mean is sensitive to outliers.
- Intuitively, then, what would outliers do to our regression estimates? It would also drag our estimate of the slope, $\hat{\beta}_1$ towards the outlier.

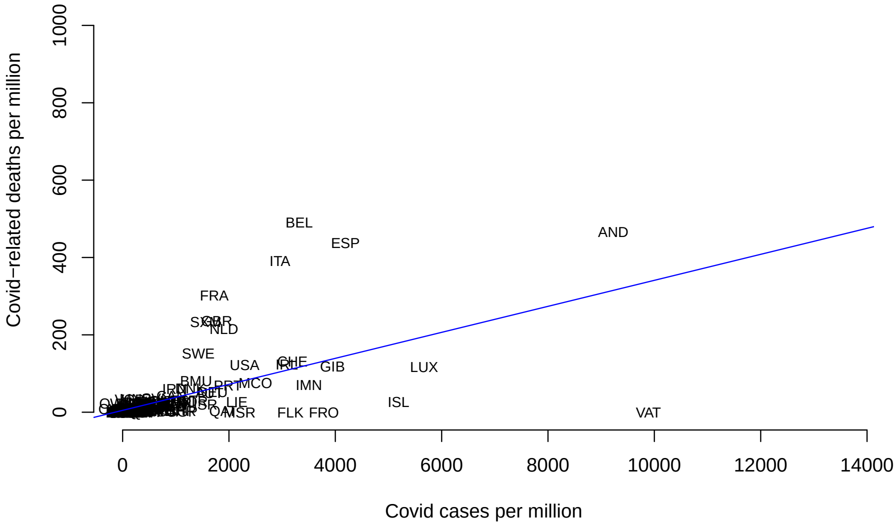
A Picture of an Outlier & Regression



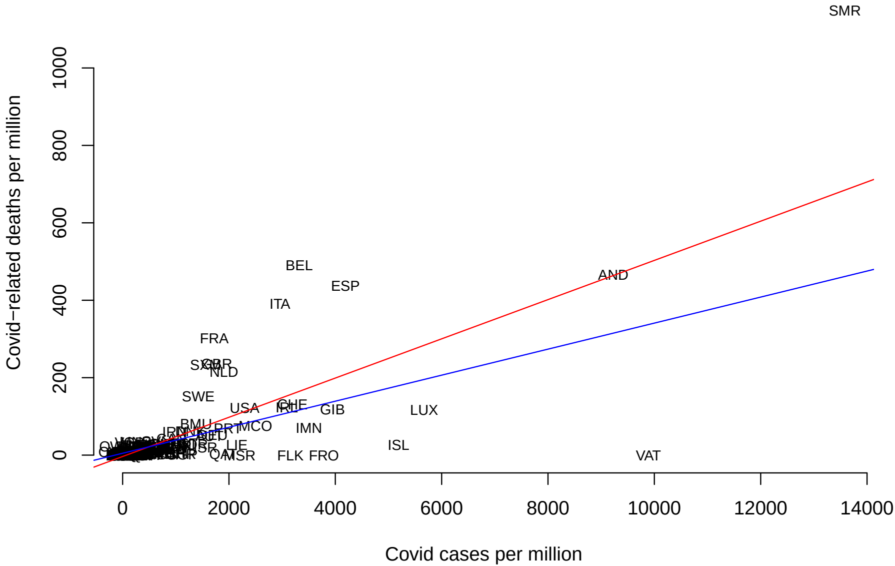
A Picture of an Outlier & Regression



A Picture of an Outlier & Regression



A Picture of an Outlier & Regression



What to do about Outliers?

- Inform readers of the issue
- Re-run the results dropping the outlier; see if this affects your results. Are they *robust* to dropping or including the outlier?
- Justify whether you should include or exclude the observation.

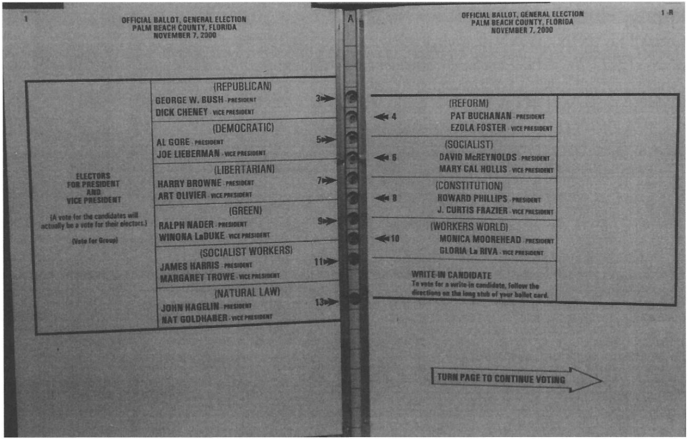
Sometimes we are interested in Outliers themselves

Is some observation really systematically different from the rest?

Wand et al. article.

What is the research question?

FIGURE 1. The Palm Beach County Bufferfly Ballot



1

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

ELECTORS
FOR PRESIDENT
AND
VICE PRESIDENT

(A vote for the candidates will actually be a vote for their electors.)

(Vote for Group)

(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3

(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5

(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7

(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9

(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT
11

(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT
13

A

1-R

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

4
(REFORM)
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT

6
(SOCIALIST)
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT

8
(CONSTITUTION)
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT

10
(WORKERS WORLD)
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT

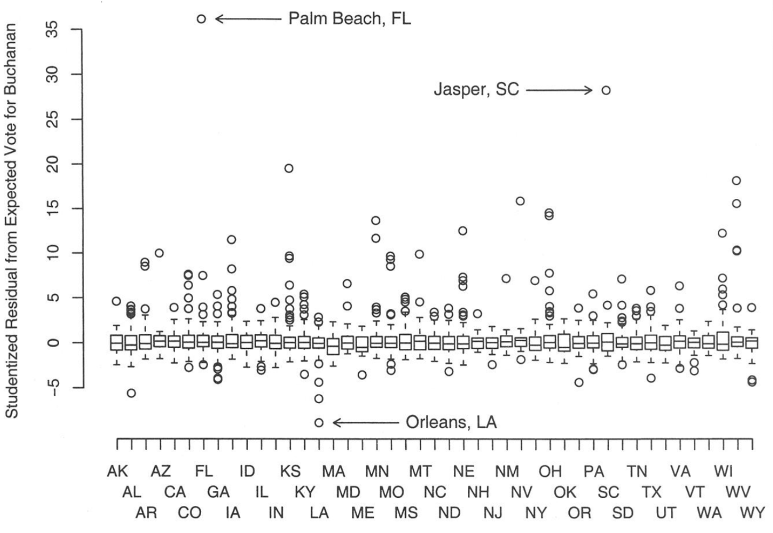
WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

TURN PAGE TO CONTINUE VOTING

Source: AP Worldwide Photos, Gary I. Rothstein. Reprinted with permission.

Sometimes we are interested in Outliers themselves

FIGURE 2. Boxplots of Studentized Residuals in U.S. Counties, by State



Bivariate OLS: Check your Knowledge

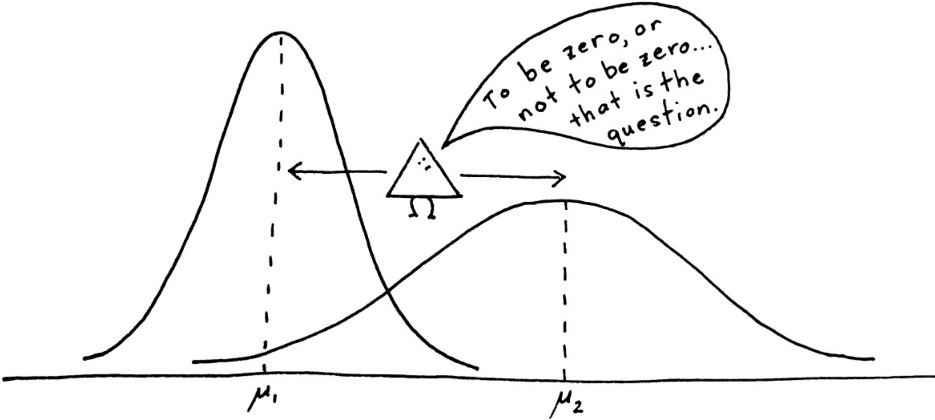
- Review these slides.

- Page 80 of the textbook (pp. 123-124 of digital version) explains what you need to know to have mastered bivariate OLS – see if you now understand all those points.
- If you've worked your way through that chapter and watched the video modules, you should be able to answer all his questions.

- Computing corner, page 83 onwards (p. 128 in digital), shows how to run a regression in R.

Next lecture: Hypothesis Testing

Where we learn how to distinguish signal from noise!
This is the pay-off from Lecture 3 on Probability.
Read Chapter 4 of Real Stats.


To be zero, or not to be zero... that is the question.
μ_1
μ_2